April 1993 LIDS-P-2173

Partial Proximal Minimization Algorithms for Convex Programming[1]

by

Dimitri P. Bertsekas[2] and Paul Tseng [3]

Abstract

We consider an extension of the proximal minimization algorithm where only some of the minimization variables appear in the quadratic proximal term. We interpret the resulting iterates in terms of the iterates of the standard algorithm and we show a uniform descent property, which holds independently of the proximal terms used. This property is used to give simple convergence proofs of parallel algorithms where multiple processors simultaneously execute proximal iterations using different partial proximal terms. We also show that partial proximal minimization algorithms are dual to multiplier methods with partial elimination of constraints, and we establish a relation between parallel proximal minimization algorithms and parallel constraint distribution algorithms.

[1] Supported by the National Science Foundation under Grant DDM-8903385 and Grant CCR-9103804, and the Army Research Office under Grant ARO DAAL03-92-G-0115.

[2] Department of Electrical Engineering and Computer Science, M. I. T., Cambridge, Mass., 02139.

[3] Department of Mathematics, Univ. of Washington, Seattle, Wash., 98195.

1. INTRODUCTION

Let us consider the proximal minimization algorithm defined by

$$x^{k+1} = \arg\min_{x \in \Re^n} \left\{ f(x) + \frac{1}{2c} \|x - x^k\|^2 \right\}. \tag{1}$$

Here c is a positive constant, $\|\cdot\|$ denotes the standard Euclidean norm on $\Re^n$, and $f : \Re^n \to (-\infty, \infty]$ is a closed proper convex function [Roc70], that is, an extended-real-valued, lower semicontinuous convex function on $\Re^n$, which is not identically $+\infty$. It is well known that, starting from an arbitrary $x^0 \in \Re^n$, the sequence $\{x^k\}$ converges to a minimizer of f if there exists at least one minimizer, and diverges otherwise.

The algorithm, originally proposed by Martinet [Mar70], [Mar72], and further refined and extended by Rockafellar [Roc76], is useful for "regularizing" the minimization of f, through the addition of the strongly convex term $\|x - x^k\|^2$. The algorithm is also useful in a dual context, where f is the dual function of a constrained nonlinear programming problem and x is a vector of Lagrange multipliers. Then, by using the Fenchel duality theorem, the proximal iteration (1) can be interpreted as an augmented Lagrangian iteration, that is, a minimization of a quadratic augmented Lagrangian function associated with the primal problem, followed by a Lagrange multiplier update. This interpretation, first given by Rockafellar [Roc73], can be found in several sources, e.g. [Ber82], [BeT89].

In this paper, we focus on a variation where the vector x is partitioned in two subvectors x_1 and x_2

$$x = (x_1, x_2)$$

with $x_1 \in \Re^{n_1}$, $x_2 \in \Re^{n_2}$, $n_1 + n_2 = n$, and the proximal term $\|x - x^k\|^2$ is replaced by the portion involving only the subvector x_1, that is,

$$(x_1^{k+1}, x_2^{k+1}) \in \arg\min_{(x_1, x_2) \in \Re^n} \left\{ f(x_1, x_2) + \frac{1}{2c} \|x_1 - x_1^k\|^2 \right\}. \tag{2}$$

We call this the *partial proximal minimization algorithm*, but hasten to observe that it can be viewed as a special case of the ordinary algorithm (1) with $f(x_1, x_2)$ replaced by

$$f_1(x_1) = \min_{x_2 \in \Re^{n_2}} f(x_1, x_2),$$

assuming that the minimum of f above is attained for all $x_2 \in \Re^{n_2}$. Thus, the sequence $\{x_1^k\}$ is in effect generated by applying the ordinary algorithm to the function f_1, while x_2^k is obtained by minimizing $f(x_1^k, x_2)$ with respect to x_2, after x_1^k has been computed. It follows that the convergence

properties of the partial algorithm can be inferred from those of the ordinary one; in fact this has been demonstrated by [Ha90]. The partial algorithm, however, allows a choice between several partial proximal terms, and also allows the simultaneous use of several different proximal terms in a parallel computing context. When such possibilities are considered, the theory of the ordinary algorithm is not directly applicable and a new analysis is needed. Our main purpose in this paper is to provide such an analysis.

Our interest in the partial algorithm stems from a recent work of Ferris and Mangasarian [FeM91] and of Ferris [Fer91] on parallel constraint distribution. Parallel constraint distribution is an augmented Lagrangian type algorithm for convex programming whereby at each iteration the constraints are partitioned into subsets and, for each subset, an augmented Lagrangian subproblem in which constraints not of the subset appear in the augmented Lagrangian is solved; the multipliers obtained from each of the subproblems are then combined in some simple fashion to yield a new set of multipliers. This algorithm has the advantages that each subproblem has fewer constraints than the original problem and the subproblems can be solved in parallel. Numerical test results reported in [FeM91] and [Fer91] indicate that the algorithm is quite promising for practical computation, especially when implemented in parallel. One of our purposes in this paper is to show that parallel constraint distribution is closely related to proximal minimization and, in particular, to a parallel implementation of the partial proximal iteration (2) (see Section 5).

A central observation of the present paper is that the partial proximal iteration (2) can be decomposed into a sequence of two steps as shown in Fig. 1:

(a) A (block) coordinate descent iteration for the function F_c defined by

$$F_c(x) = \min_{y \in \Re^n} \left\{ f(y) + \frac{1}{2c}\|y - x\|^2 \right\}.$$

This iteration is done with respect to the second coordinate subvector x_2 of the vector $x = (x_1, x_2)$; it starts at the current vector (x_1^k, x_2^k), and yields a vector (x_1^k, x_2^{k+1}).

(b) An iteration of the ordinary proximal algorithm starting at the vector (x_1^k, x_2^{k+1}) obtained from the preceding coordinate descent iteration; this can be interpreted as a gradient iteration with stepsize equal to c for minimizing the same function F_c (see e.g. [BeT89, p. 234]).

By contrast, the ordinary proximal iteration does only step (b) above. Thus the partial iteration differs from the ordinary one only in that it executes an extra coordinate descent step prior to each ordinary proximal iteration. Note here that F_c is continuously differentiable and has the same minimizers and minimum value as f; see e.g. [BeT89]. Thus both steps (a) and (b) above are aimed at approaching an optimal solution.

A consequence of the preceding observation is that the value of the function F_c provides a uniform

criterion of merit, which is improved by all partial proximal iterations, independently of the partition (x_1, x_2). We use this fact in Section 3 to provide a short convergence proof for a parallel algorithm that involves execution of different partial proximal iterations by different processors. In Section 4, we derive the rate of convergence of this algorithm. In Section 5, we show that partial proximal minimization algorithms are intimately related to augmented Lagrangian algorithms with partial elimination of constraints. When specialized within the augmented Lagrangian context, the parallel algorithm of Section 3 becomes similar to the parallel constraint distribution algorithms of Ferris and Mangasarian [FeM91], [Fer91a]; however, our convergence proofs are less complicated than those in [FeM91] and [Fer91a], and require less restrictive assumptions. In particular, the convergence analysis of [FeM91] and [Fer91a] assumes that the cost function is positive definite quadratic and the constraints are linear, while we assume general convex cost and constraint functions. Most of our analysis carries through to partial proximal minimization algorithms with nonquadratic proximal terms [Ber82], [CeZ92], [ChT90], [Eck93], [KoB76], [GoT79], [Luq84], [Luq86], [TsB90]. We thus take these more general methods as our starting point and specialize our results to the case of quadratic proximal terms whenever the results for this case are stronger. In particular, our algorithms and corresponding convergence results are patterned after those of Kort and Bertsekas [KoB76]. On the other hand, our analysis assumes that the proximal term contains the origin in its interior, and thus does not apply to methods using logarithmic/entropy proximal terms [Ber82], [CeZ92], [ChT90], [Eck93], [TsB90], and the corresponding augmented Lagrangian methods that use the exponential penalty function [KoB72], [Ber82], [TsB90].

2. A UNIFORM DESCENT PROPERTY

In this section we introduce the notion of partial proximal minimization and analyze its descent properties. These descent properties will be used later to establish the convergence of algorithms based on successive applications of partial proximal minimization.

We first define partial proximal minimization in the general context of nonquadratic proximal terms. Consider the class of strictly convex, continuously differentiable functions $\phi : \Re \to \Re$ such that

$$\phi(0) = 0, \qquad \nabla\phi(0) = 0, \qquad \lim_{t\to-\infty} \nabla\phi(t) = -\infty, \qquad \lim_{t\to\infty} \nabla\phi(t) = \infty.$$

This class was introduced in [KoB76] within the dual context of nonquadratic augmented Lagrangian methods, together with the generalization of the proximal minimization algorithm obtained by replacing "$|\cdot|^2/2$" with "$\phi(\cdot)$". For an extensive discussion of the subject, see [Ber82, Ch. 5]. A

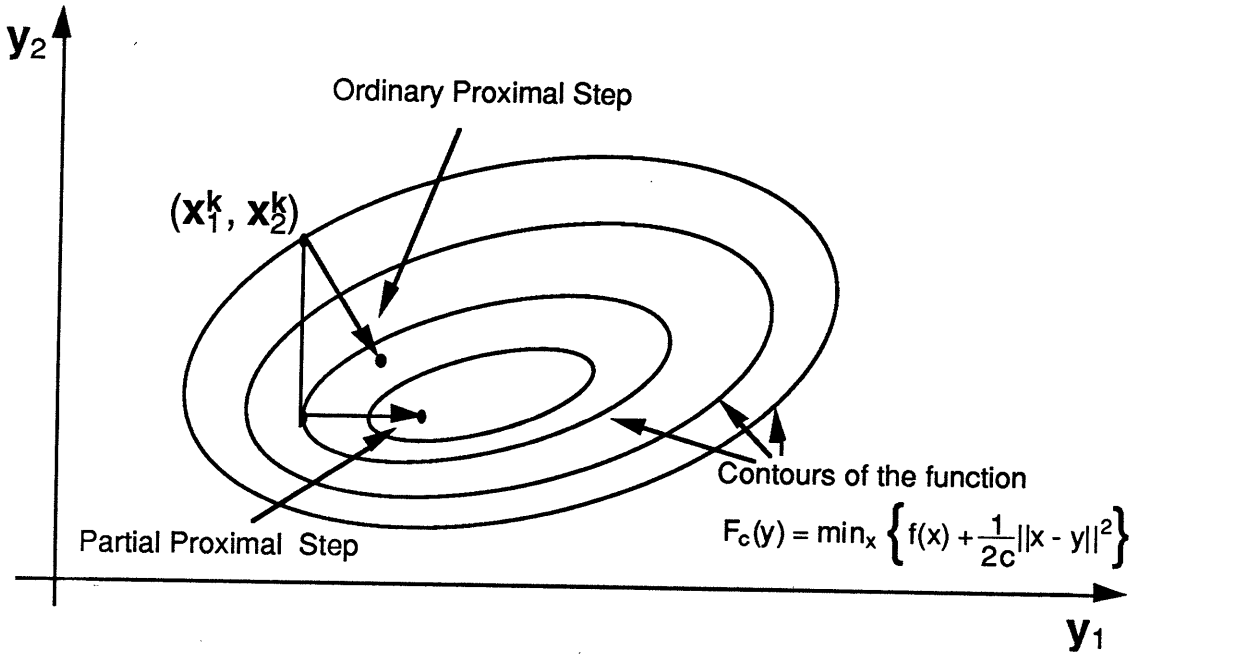

Figure 1: An illustration of the ordinary and the partial proximal iterations starting at the vector $x^k = (x_1^k, x_2^k)$. Both iterations involve a gradient step on the function F_c with stepsize c. However, the partial proximal iteration precedes the gradient step with a coordinate descent step along the subvector x_2.

prominent example in the class is the power function: $\phi(t) = (1/\gamma)|t|^\gamma$ with $\gamma > 1$. For $\gamma = 2$ we obtain the quadratic function used earlier.

For each $c > 0$, let F_c be the real-valued convex function on $\Re^n$ defined by

$$F_c(x) = \min_{y \in \Re^n} \left\{ f(y) + \frac{1}{c}\Phi(y - x) \right\}, \tag{3}$$

where $\Phi : \Re^n \mapsto \Re$ is the function

$$\Phi(t_1, ..., t_n) = \sum_{i=1}^{n} \phi(t_i).$$

We remark that we can have different ϕ's for different coordinate indices i but, for simplicity, we will not consider this more general case. We also note that the gradient mapping $\nabla\Phi : \Re^n \to \Re^n$ has an inverse $\nabla\Phi^{-1} : \Re^n \to \Re^n$ because of the defining properties of the function Φ. We have the following proposition which formalizes the interpretation of the partial proximal iteration as a block coordinate descent step followed by an ordinary proximal step (compare with Fig. 1).

Proposition 1: Let $c > 0$ and a subset I of the index set $\{1, \ldots, n\}$ be given. For any $x \in \Re^n$, consider a vector x' satisfying

$$x' \in \arg\min_{y \in \Re^n} \left\{ f(y) + \frac{1}{c} \sum_{i \in I} \phi(y_i - x_i) \right\} \tag{4}$$

and let x'' be the vector with components

$$x_i'' = \begin{cases} x_i & \forall\, i \in I \\ x_i' & \forall\, i \notin I. \end{cases} \tag{5}$$

Then

$$x' = x'' + \nabla\Phi^{-1}\left(-c\nabla F_c(x'')\right) = \arg\min_{y\in\Re^n}\left\{f(y) + \frac{1}{c}\Phi(y - x'')\right\}, \tag{6}$$

$$x'' \in \arg\min_{\{y\,|\,y_i = x_i,\ i\in I\}} F_c(y), \tag{7}$$

where $F_c : \Re^n \to \Re$ is the convex function defined by (3). Conversely if x' and x'' satisfy (6) and (7), then they also satisfy (4) and (5).

Proof: Fix any $x \in \Re^n$. Let x' be a vector satisfying (4) and let x'' be given by (5). We will show that (6) and (7) hold. Indeed, from the definition of x'', the vector x' minimizes over $y \in \Re^n$ not only $f(y) + \frac{1}{2c}\sum_{i\in I}\phi(y_i - x_i)$ but also $\frac{1}{2c}\sum_{i\notin I}\phi(y_i - x_i'')$, implying that x' minimizes the sum, which is $f(y) + \frac{1}{2c}\Phi(y - x'')$. This proves the second equality in (6). The first equality in (6) is due to the invertibility of $\nabla\Phi$ and the following consequence of Prop. 5.5(c) in [Ber82]:

$$\nabla\Phi\big(x' - x''\big) = -c\nabla F_c(x'').$$

[Actually Prop. 5.5(c) of [Ber82] addresses the dual context of nonquadratic augmented Lagrangian methods, and thus considers (dual) functions f with $f(x) = \infty$ for all x outside the nonnegative orthant. The proof given in [Ber82], however, applies verbatim to the more general convex function f considered here.]

To prove (7), note that for all vectors $z \in \Re^n$ with $z_i = x_i$ for all $i \in I$, we have

$$\begin{aligned} F_c(z) &= \min_{y\in\Re^n}\left\{f(y) + \frac{1}{2c}\sum_{i\in I}\phi(y_i - x_i) + \frac{1}{2c}\sum_{i\notin I}\phi(y_i - z_i)\right\} \\ &\geq \min_{y\in\Re^n}\left\{f(y) + \frac{1}{2c}\sum_{i\in I}\phi(y_i - x_i)\right\} \\ &= f(x') + \frac{1}{2c}\sum_{i\in I}\phi(x_i' - x_i) \\ &= f(x') + \frac{1}{2c}\sum_{i\in I}\phi(x_i' - x_i) + \frac{1}{2c}\sum_{i\notin I}\phi(x_i' - x_i') \\ &\geq F_c(x''), \end{aligned}$$

where the last inequality follows from the definitions of x'' and F_c. This proves (7).

Conversely, suppose that x' and x'' satisfy (6) and (7). We will show that (4) and (5) hold. Indeed, (7) implies that $x_i = x_i''$ for all $i \in I$, and that $\partial F_c(x'')/\partial x_i = 0$ for all $i \notin I$, so from (6) we

have $x_i' = x_i''$ for all $i \notin I$. Thus (5) holds. To show (4) we argue by contradiction. Suppose that for some $z \in \Re^n$ we have

$$f(z) + \frac{1}{2c}\sum_{i\in I}\phi(z_i - x_i) < f(x') + \frac{1}{2c}\sum_{i\in I}\phi(x_i' - x_i).$$

Then the directional derivative of the function $y \mapsto f(y) + \frac{1}{2c}\sum_{i\in I}\phi(y_i - x_i)$ at x' along the direction $z - x'$ is negative. This directional derivative is equal to the directional derivative of the function $y \mapsto f(y) + \frac{1}{2c}\sum_{i\in I}\phi(y_i - x_i) + \frac{1}{2c}\sum_{i\notin I}\phi(y_i - x_i')$ at x' along the direction $z - x'$. The latter directional derivative, however, is nonnegative in view of (5) and (6), arriving at a contradiction. This proves (4). **Q.E.D.**

By using Prop. 1 and the special structure of F_c, we obtain the following key descent property for F_c under the partial proximal iteration. This property will be used in the subsequent convergence analysis.

Lemma 1: Let $c > 0$ and a subset I of $\{1, \ldots, n\}$ be given. Let $F_c : \Re^n \to \Re$ be the continuously differentiable convex function given by (3).

(a) For any $x \in \Re^n$, the vector x' given by (4) satisfies

$$F_c(x') - F_c(x'') \leq -\frac{1}{c}\Phi(x' - x''),$$

where x'' is the vector given by (5).

(b) Let $\phi(\cdot) = \frac{1}{2}|\cdot|^2$ [so that $\Phi(\cdot)$ is the quadratic function $\frac{1}{2}\|\cdot\|^2$]. Then, for any $x \in \Re^n$, the vector x' given by (4) satisfies

$$F_c(x') - F_c(x) \leq -\frac{c}{4}\|\nabla F_c(x)\|^2.$$

Proof: (a) Fix any $x \in \Re^n$ and let x' and x'' be given by (4) and (5), respectively. By Prop. 1, x' and x'' satisfy (6) and so the definition of F_c yields

$$F_c(x'') = f(x') + \frac{1}{c}\Phi(x' - x'').$$

It is also easily seen from the definition of F_c that $F_c(x') \leq f(x')$, which together with the above equality implies the result.

(b) We first establish some basic inequalities satisfied by the function F_c and its gradient. For any $x, y \in \Re^n$, let

$$\tilde{x} = \arg\min_{z\in\Re^n}\left\{f(z) + \frac{1}{2c}\|z - x\|^2\right\}, \qquad \tilde{y} = \arg\min_{z\in\Re^n}\left\{f(z) + \frac{1}{2c}\|z - y\|^2\right\}.$$

Then,

$$\frac{1}{c}(y-\tilde{y}) \in \partial f(\tilde{y}),$$

and moreover, using (6) and $\Phi(\cdot) = \frac{1}{2}||\cdot||^2$,

$$x - \tilde{x} = c\nabla F_c(x), \qquad y - \tilde{y} = c\nabla F_c(y).$$

Combining these three relations and using the convexity of f, we obtain

$$\begin{aligned} F_c(x) - F_c(y) - \langle \nabla F_c(y), x-y\rangle &= f(\tilde{x}) + \frac{1}{2c}||\tilde{x}-x||^2 - f(\tilde{y}) - \frac{1}{2c}||\tilde{y}-y||^2 - \frac{1}{c}\langle y-\tilde{y}, x-y\rangle \\ &= f(\tilde{x}) - f(\tilde{y}) - \frac{1}{c}\langle y-\tilde{y}, \tilde{x}-\tilde{y}\rangle + \frac{1}{2c}||\tilde{x}-x-(\tilde{y}-y)||^2 \\ &\geq \frac{1}{2c}||\tilde{x}-x-(\tilde{y}-y)||^2 \\ &= \frac{c}{2}||\nabla F_c(x) - \nabla F_c(y)||^2 \qquad \forall\ x, y \in \Re^n. \end{aligned} \tag{8}$$

Let us now fix $x \in \Re^n$, and let x' and x'' be given by (4) and (5), respectively. By Prop. 1, x' and x'' satisfy (6), so the assumption $\Phi(\cdot) = \frac{1}{2}||\cdot||^2$ implies $x' = x'' - c\nabla F_c(x'')$. Then part (a) yields

$$F_c(x') - F_c(x'') \leq -\frac{1}{2c}||x'-x''||^2 = -\frac{c}{2}||\nabla F_c(x'')||^2.$$

Relation (7) implies $\langle \nabla F_c(x''), x - x''\rangle = 0$, so by invoking (8), we also obtain

$$\frac{c}{2}||\nabla F_c(x) - \nabla F_c(x'')||^2 \leq F_c(x) - F_c(x'').$$

Adding the above two relations and rearranging terms yield

$$\frac{c}{2}\big(||\nabla F_c(x) - \nabla F_c(x'')||^2 + ||\nabla F_c(x'')||^2\big) \leq F_c(x) - F_c(x'),$$

so, by using the following easily verifiable inequality

$$||\nabla F_c(x)||^2 \leq 2\big(||\nabla F_c(x) - \nabla F_c(x'')||^2 + ||\nabla F_c(x'')||^2\big),$$

the result follows. **Q.E.D.**

The following proposition provides some additional inequalities, which compare the iterates of the ordinary and the partial proximal algorithms, and are useful for the convergence analysis of the latter (see the proof of Prop. 2).

Lemma 2: Let $c > 0$ and a subset I of $\{1, \ldots, n\}$ be given. For any $x \in \Re^n$, the vector x' given by (4) satisfies

$$f(x') \leq F_c(x) \leq f(x),$$

where $F_c : \Re^n \to \Re$ is the convex function defined by (3).

Proof: We have

$$\begin{aligned} f(x') &\leq f(x') + \frac{1}{2c}\sum_{i\in I}\phi(x'_i - x_i) \\ &= \min_{y\in\Re^n}\left\{f(y) + \frac{1}{2c}\sum_{i\in I}\phi(y_i - x_i)\right\} \\ &\leq \min_{y\in\Re^n}\left\{f(y) + \frac{1}{2c}\Phi(y - x)\right\} \\ &\leq f(x). \end{aligned}$$

Since the expression in the right-hand side of the second inequality is equal to $F_c(x)$, the result follows. **Q.E.D.**

Note an important consequence of Lemma 2: if f is bounded below and if $\{x^k\}$ is a sequence generated by the partial proximal minimization iteration (2), then both $\{f(x^k)\}$ and $\{F_c(x^k)\}$ converge monotonically to the same value, regardless of the particular partition used. It is possible to change the partition from one iteration to the next if this can improve convergence. Furthermore, if ϕ is a quadratic function, we have from Lemma 1(b) that $\nabla F_c(x^k) \to 0$, so that all cluster points of $\{x^k\}$ minimize F_c and hence also f. This result will be extended in the next section when we consider parallel versions of the partial proximal minimization algorithm.

3. PARALLEL ALGORITHMS

We now consider the following extension of the partial proximal minimization algorithm. At the start of the kth iteration we have the current iterate x^k. We construct all distinct partitions of x^k into two subvectors, and we execute the partial proximal iteration corresponding to each partition and to a chosen scalar c^k. The next iterate x^{k+1} is an arbitrary convex combination of the different vectors thus obtained. We describe below this algorithm, which we call the *parallel partial proximal minimization algorithm* (or the parallel PPM algorithm, for short).

Let $\mathcal{C}$ denote the set of all subsets of $\{1,\ldots,n\}$. Beginning with an arbitrary $x^0 \in \Re^n$, we generate a sequence $\{x^k\}$ as follows: Given x^k, we choose a scalar $c^k > 0$ and, for every subset $I \in \mathcal{C}$, let $\tilde{x}_I^k$ be given by

$$\tilde{x}_I^k \in \arg\min_{y\in\Re^n}\left\{f(y) + \frac{1}{c^k}\sum_{i\in I}\phi(y_i - x_i^k)\right\}; \tag{9}$$

we define x^{k+1} to be an arbitrary convex combination of the vectors $\tilde{x}_I^k$, that is,

$$x^{k+1} = \sum_{I \in \mathcal{C}} \alpha_I^k \tilde{x}_I^k, \tag{10}$$

where α_I^k, $I \in \mathcal{C}$, are any scalars satisfying

$$\sum_{I \in \mathcal{C}} \alpha_I^k = 1, \qquad \alpha_I^k \geq 0 \quad \forall\, I \in \mathcal{C}. \tag{11}$$

(Note that an α_I^k may be zero, so we need only compute the vectors $\tilde{x}_I^k$ with $\alpha_I^k > 0$.)

In order for the parallel PPM algorithm to be well defined, we assume that the minimum in each partial proximal iteration [cf. (9)] is attained. Note that the multiple partial proximal iterations involved in (10)-(11) can be executed in parallel by multiple processors. The next proposition shows the validity of the algorithm by combining the inequalities of Lemmas 1 and 2, and by using a modification of the convergence arguments for the ordinary proximal minimization algorithm (see [BeT89, p. 240]).

Proposition 2: Let $\{x^k\}$ be a sequence generated by the parallel PPM algorithm (9)-(11) with $\{c^k\}$ monotonically nondecreasing.

(a) If the set of minimizers of f is nonempty and compact, then $\{x^k\}$ is bounded, each of its cluster points is a minimizer of f, and $\lim_{k\to\infty} f(x^k) = \min_x f(x)$.

(b) Assume that f is bounded below and let $\phi(\cdot) = \frac{1}{2}|\cdot|^2$ [so $\Phi(\cdot)$ is the quadratic function $\frac{1}{2}\|\cdot\|^2$]. Both $\{f(x^k)\}$ and $\{F_{c^k}(x^k)\}$ are monotonically nonincreasing and $\lim_{k\to\infty} \nabla F_{c^k}(x^k) = 0$, where $F_c : \Re^n \to \Re$ is the function given by (3).

Proof: (a) For each k and $I \in \mathcal{C}$, we have by applying Lemma 2 with $x = x^k$, $c = c^k$, and $x' = \tilde{x}_I^k$ that

$$f(\tilde{x}_I^k) \leq F_{c^k}(x^k) \leq f(x^k).$$

This together with (10)-(11) and the convexity of f yields

$$f(x^{k+1}) \leq \sum_{I \in \mathcal{C}} \alpha_I^k f(\tilde{x}_I^k) \leq f(x^k).$$

Thus, $\{f(x^k)\}$ is monotonically nonincreasing.

For each k and $I \in \mathcal{C}$, since $\tilde{x}_I^k$ is given by (9), we have from Prop. 1 that

$$\tilde{x}_I^k = \arg\min_{y \in \Re^n} \left\{ f(y) + \frac{1}{c^k} \Phi(y - \hat{x}_I^k) \right\}, \tag{12}$$

where $\hat{x}_I^k$ is the vector in $\Re^n$ whose ith component is the ith component of x^k if $i \in I$ and, otherwise, is the ith component of $\tilde{x}_I^k$. We also have from (7) that, for each k and $I \in \mathcal{C}$,

$$F_{c^k}(\hat{x}_I^k) \leq F_{c^k}(x^k),$$

and from Lemma 1(a) that

$$F_{c^k}(\tilde{x}_I^k) \leq F_{c^k}(\hat{x}_I^k) - \frac{1}{c^k}\Phi(\tilde{x}_I^k - \hat{x}_I^k).$$

Combining the above two relations and using (10)-(11) and the convexity of F_{c^k} yields

$$F_{c^k}(x^{k+1}) \leq F_{c^k}(x^k) - \frac{1}{c^k}\sum_{I\in\mathcal{C}} \alpha_I^k \Phi\big(\tilde{x}_I^k - \hat{x}_I^k\big).$$

Since $F_{c^{k+1}}(x^{k+1}) \leq F_{c^k}(x^{k+1})$ [cf. $c^{k+1} \geq c^k$ and (3)], this shows that $\{F_{c^k}(x^k)\}$ is monotonically nonincreasing and that

$$\lim_{k\to\infty} \frac{1}{c^k}\sum_{I\in\mathcal{C}} \alpha_I^k \Phi\big(\tilde{x}_I^k - \hat{x}_I^k\big) = 0. \tag{13}$$

Also, the compactness of the set of minimizers of f implies that all the level sets of f are compact [Roc70, Cor. 8.7.1] and that, for $k \geq 1$ and $I \in \mathcal{C}$, x^k and $\tilde{x}_I^k$ are contained in the level set $\{x \mid f(x) \leq f(x^0)\}$ (cf. Lemma 2). It follows that the sequences $\{x^k\}$ and $\{\tilde{x}_I^k\}$ are bounded. Since each component of $\hat{x}_I^k$ is either a component of x^k or of $\tilde{x}_I^k$, it follows that the sequence $\{\hat{x}_I^k\}$ is also bounded.

Fix any minimizer x^* of f. Using (12) and the convexity of f, we have for each $\xi \in (0,1)$ and $I \in \mathcal{C}$ that

$$\begin{aligned} f(\tilde{x}_I^k) + \frac{1}{c^k}\Phi(\tilde{x}_I^k - \hat{x}_I^k) &\leq f\big(\xi x^* + (1-\xi)\tilde{x}_I^k\big) + \frac{1}{c^k}\Phi\big(\tilde{x}_I^k - \hat{x}_I^k + \xi(x^* - \tilde{x}_I^k)\big) \\ &\leq \xi f(x^*) + (1-\xi) f(\tilde{x}_I^k) + \frac{1}{c^k}\Phi\big(\tilde{x}_I^k - \hat{x}_I^k + \xi(x^* - \tilde{x}_I^k)\big). \end{aligned}$$

Rearranging terms and dividing both sides by ξ gives

$$f(\tilde{x}_I^k) \leq f(x^*) + \frac{1}{\xi c^k}\left[\Phi\big(\tilde{x}_I^k - \hat{x}_I^k + \xi(x^* - \tilde{x}_I^k)\big) - \Phi(\tilde{x}_I^k - \hat{x}_I^k)\right],$$

which together with (10)-(11) and the convexity of f yields

$$\begin{aligned} f(x^{k+1}) &\leq f(x^*) + \frac{1}{\xi c^k}\sum_{I\in\mathcal{C}} \alpha_I^k \left[\Phi\big(\tilde{x}_I^k - \hat{x}_I^k + \xi(x^* - \tilde{x}_I^k)\big) - \Phi(\tilde{x}_I^k - \hat{x}_I^k)\right] \\ &\leq f(x^*) + \frac{1}{c^k}\sum_{I\in\mathcal{C}} \alpha_I^k \langle \nabla\Phi\big(\tilde{x}_I^k - \hat{x}_I^k + \xi(x^* - \tilde{x}_I^k)\big), x^* - \tilde{x}_I^k\rangle, \end{aligned}$$

where the second inequality follows from the convexity of Φ. By taking the limit supremum as $k \to 0$, we obtain

$$\limsup_{k\to\infty} f(x^{k+1}) \leq f(x^*) + \limsup_{k\to\infty} \frac{1}{c^k}\sum_{I\in\mathcal{C}} \alpha_I^k \langle \nabla\Phi\big(\tilde{x}_I^k - \hat{x}_I^k + \xi(x^* - \tilde{x}_I^k)\big), x^* - \tilde{x}_I^k\rangle.$$

Finally, we take the limit of both sides as $\xi \to 0$. Since the sequences $\{\tilde{x}_I^k\}$, $\{\hat{x}_I^k\}$ and $\{1/c^k\}$ are all bounded, we can pass this limit through the limit supremum on the right-hand side to obtain

$$\limsup_{k\to\infty} f(x^{k+1}) \leq f(x^*) + \limsup_{k\to\infty} \frac{1}{c^k}\sum_{I\in\mathcal{C}} \alpha_I^k \langle \nabla\Phi\big(\tilde{x}_I^k - \hat{x}_I^k\big), x^* - \tilde{x}_I^k\rangle. \tag{14}$$

Assume first that $c^k \to \infty$. Since both sequences $\{\tilde{x}_I^k\}$ and $\{\hat{x}_I^k\}$ are bounded, it follows from (14) that $\limsup_{k\to\infty} f(x^{k+1}) \leq f(x^*)$. Since x^* minimizes f, $\{f(x^k)\}$ must converge to $f(x^*)$.

Assume now that $c^k \to \bar{c} < \infty$. Let $\mathcal{K}$ be an infinite subsequence of the set of positive integers such that for every $I \in \mathcal{C}$, $\{\alpha_I^k\}_{k\in\mathcal{K}}$ converges (to either a positive number or zero) with $\lim_{k\to\infty,\ k\in\mathcal{K}} \alpha_I^k > 0$ for at least one $I \in \mathcal{C}$; since $\mathcal{C}$ is a finite set, such a subsequence exists. From (13) we see that for all $I \in \mathcal{C}$ such that $\lim_{k\to\infty,\ k\in\mathcal{K}} \alpha_I^k > 0$, we have

$$\lim_{k\to\infty,\ k\in\mathcal{K}} \Phi(\tilde{x}_I^k - \hat{x}_I^k) = 0$$

and hence also $\lim_{k\to\infty,\ k\in\mathcal{K}} \{\tilde{x}_I^k - \hat{x}_I^k\} = 0$, implying that

$$\lim_{k\to\infty,\ k\in\mathcal{K}} \nabla\Phi(\tilde{x}_I^k - \hat{x}_I^k) = 0.$$

Using the above equation in (14), we obtain $\limsup_{k\to\infty,\ k\in\mathcal{K}} f(x^{k+1}) \leq f(x^*)$. Since x^* minimizes f, $\{f(x^k)\}_{k\in\mathcal{K}}$ must converge to $f(x^*)$. Since $\{f(x^k)\}$ is monotonically nonincreasing, it too must converge to $f(x^*)$.

(b) Fix any k. For each $I \in \mathcal{C}$, we have upon applying Lemma 1(b) with $x = x^k$, $c = c^k$, and x' equal to the vector $\tilde{x}_I^k$ given by (9) that

$$F_{c^k}(\tilde{x}_I^k) - F_{c^k}(x^k) \leq -\frac{c^k}{4}||\nabla F_{c^k}(x^k)||^2.$$

Combining this with (10)-(11) and the convexity of F_{c^k}, we obtain

$$\begin{aligned} F_{c^k}(x^{k+1}) - F_{c^k}(x^k) &\leq \sum_{I\in\mathcal{C}} \alpha_I^k \big(F_{c^k}(\tilde{x}_I^k) - F_{c^k}(x^k)\big) \\ &\leq -\frac{c^k}{4} \sum_{I\in\mathcal{C}} \alpha_I^k ||\nabla F_{c^k}(x^k)||^2 \\ &= -\frac{c^k}{4}||\nabla F_{c^k}(x^k)||^2. \end{aligned}$$

Since $F_{c^{k+1}}(x^{k+1}) \leq F_{c^k}(x^{k+1})$ [cf. $c^{k+1} \geq c^k$ and (3)], and F_{c^k} is bounded below by $\inf_{x\in\Re^n} f(x)$ for every k, the preceding relation implies that $\{F_{c^k}(x^k)\}$ is monotonically nonincreasing and that

$$\lim_{k\to\infty} c^k ||\nabla F_{c^k}(x^k)||^2 = 0.$$

Since $\{c^k\}$ is bounded below by $c^0 > 0$, this proves that $\lim_{k\to\infty} \nabla F_{c^k}(x^k) = 0$.

For each $I \in \mathcal{C}$, we have upon applying Lemma 2 with $x = x^k$, $c = c^k$, and [cf. (9)] $x' = \tilde{x}_I^k$ that

$$f(\tilde{x}_I^k) \leq F_{c^k}(x^k) \leq f(x^k).$$

This together with (10)-(11) and the convexity of f yields

$$f(x^{k+1}) \leq \sum_{I\in\mathcal{C}} \alpha_I^k f(\tilde{x}_I^k) \leq F_{c^k}(x^k) \leq f(x^k).$$

Thus, $\{f(x^k)\}$ is monotonically nonincreasing. **Q.E.D.**

It can be seen from the above proof that the monotonicity property of $\{c^k\}$ is not crucial for Proposition 2 to hold. Instead, it suffices that $\{F_{c^k}(x^k)\}$ is maintained nonincreasing and $\{c^k\}$ is bounded away from zero.

The parallel PPM algorithm can be generalized, and can be implemented in a flexible and asynchronous manner. One can envision multiple processors executing asynchronously different partial proximal iterations starting from the vector that is currently best in terms of some uniform merit criterion, such as the value of f or F_c. The results from several processors can be combined via a convex combination, but the convex combination should be accepted only if the uniform merit criterion is improved.

It should be noted that the above convergence results for the parallel PPM algorithm are considerably weaker than those available for the proximal minimization algorithm (see [BeT89, Sec. 3.4.3], [ChT90], [EcB92], [Fer91b], [GoT79], [Gul91], [Luq86], [Mar70], [Roc76]). One difficulty is that the extra coordinate descent step of (7) destroys certain monotonicity properties of the iterates, which are essential to proving some of the stronger convergence properties of the proximal minimization algorithm.

Finally, we note that the choice of the coefficients α_I^k, $I \in \mathcal{C}$, can have a significant effect on the convergence rate of the parallel PPM algorithm. To illustrate, consider applying the parallel PPM algorithm with $\phi(\cdot) = \frac{1}{2}|\cdot|^2$ to minimize the two-dimensional convex differentiable function

$$f(x_1, x_2) = \frac{1}{2}\big\{\max\{0, x_1\}^2 + \max\{0, x_2\}^2\big\}.$$

Suppose furthermore that $x_1^0 > 0$ and $x_2^0 > 0$, and that $c^k = c > 0$ for all k. If we set $\alpha_{\{1\}}^k = 1$ (so $\alpha_{\{2\}}^k = \alpha_{\{1,2\}}^k = 0$) for all k, then it is not difficult to see that one possible sequence is given by $x_1^k = x_1^0/(1+c)^k$ and $x_2^k = (-1)^k - 1$ for all k, so the cost converges at a linear rate but the iterates themselves do not converge. On the other hand, if we set $\alpha_{\{1\}}^k$ to alternate between 1 and 0 with the corresponding values of $\alpha_{\{2\}}^k$ alternating between 0 and 1, while $\alpha_{\{1,2\}}^k = 0$, then it can be seen that a minimizer of f is obtained after only two iterations.

4. RATE OF CONVERGENCE

We now turn to the analysis of the convergence rate of the parallel PPM algorithm. To establish some terminology, consider a real sequence $\{s_k\}$ that converges to a real number s^*. We say that

$\{s_k\}$ converges *finitely* if there exists $\overline{k}$ such that $s_k = s^*$ for all $k \geq \overline{k}$; *superlinearly with order p*, where $p > 1$, if

$$\limsup_{k\to\infty} \frac{|s_{k+1} - s^*|}{|s_k - s^*|^p} < \infty;$$

superlinearly if

$$\lim_{k\to\infty} \frac{|s_{k+1} - s^*|}{|s_k - s^*|} = 0;$$

and *linearly* if there exist $\beta \in (0,1)$, and $\overline{k}$ such that

$$\frac{|s_{k+1} - s^*|}{|s_k - s^*|} \leq \beta \qquad \forall\ k \geq \overline{k}.$$

Following [OrR70], we say that $\{s_k\}$ converges *R-linearly* to s^* if there exist scalars $q > 0$ and $\beta \in [0,1)$ such that $|s_k - s^*| \leq q\beta^k$ for all k. Note that $\{s_k\}$ converges R-linearly if $|s_k - s^*| \leq t_k$ for all k where $\{t_k\}$ is some sequence converging linearly to zero.

The convergence rate of the ordinary proximal minimization algorithm depends on the growth properties of the minimized function f as well as the growth properties of the proximal term used. The following key assumption was first introduced in [KoB76] (see also [Ber82, p. 342]) and was used to analyze the convergence rate of the proximal minimization algorithm for quadratic as well as certain types of nonquadratic proximal terms.

Assumption A:

The set of minimizers of f, denoted X^*, is nonempty and compact. Furthermore, there exist scalars $\alpha \geq 1$, $\beta > 0$, and $\delta > 0$ such that

$$\beta\big(\rho(x;X^*)\big)^\alpha \leq f(x) - \min_{y\in\Re^n} f(y) \qquad \forall\ x \text{ with } \rho(x;X^*) \leq \delta, \tag{15}$$

where $\rho(x;X^*)$ is the distance from x to X^* given by

$$\rho(x;X^*) = \min_{x^*\in X^*} ||x - x^*||. \tag{16}$$

The ordinary proximal minimization algorithm has finite, superlinear, or linear convergence rate depending on whether $\alpha = 1$, $1 < \alpha < 2$, or $\alpha = 2$, respectively; see references [KoB76], [Ber75] (which deals with the case $\alpha = 1$), and [Ber82], Chapter 5 (which provides a comprehensive analysis). The convergence rate is also superlinear if $\alpha = 2$ and $c^k \to \infty$. If $\alpha > 2$, the convergence rate is slower than linear, that is, some of the generated sequences do not converge linearly. In the case where the proximal term has a growth rate $\gamma > 1$ other than quadratic ($\gamma \neq 2$), the convergence rate is influenced by γ (it is superlinear if $1 < \alpha < \gamma$ even in the case where $\alpha \geq 2$).

The following proposition provides corresponding, although slightly weaker, results for the parallel PPM algorithm.

Proposition 3: Let Assumption A hold, let $f^* = \min_{y \in \Re^n} f(y)$, and let $\{x^k\}$ be a sequence generated by the parallel PPM algorithm (9)-(11) with $\{c^k\}$ monotonically nondecreasing.

(a) If $\alpha = 1$ and there exists a scalar $\overline{\alpha} > 0$ such that

$$\alpha_I^k \geq \overline{\alpha} \qquad \forall\ k \text{ and } I \in \mathcal{C} \text{ such that } \alpha_I^k > 0, \tag{17}$$

then $\{f(x^k)\}$ converges to f^* finitely.

(b) Assume that for some scalars $M > 0$ and $\gamma \geq \alpha$ we have

$$\Phi(x) \leq M||x||^\gamma \qquad \forall\ x \text{ with } ||x|| \leq \delta. \tag{18}$$

If $1 < \alpha < \gamma$, then $\{f(x^k)\}$ converges to f^* superlinearly with order γ/α. Also, if $1 < \alpha = \gamma$ and $c^k \to \infty$, then $\{f(x^k)\}$ converges to f^* superlinearly.

(c) Let $\phi(\cdot) = \frac{1}{2}|\cdot|^2$ [so $\Phi(\cdot)$ is the quadratic function $\frac{1}{2}||\cdot||^2$]. If $\alpha = 2$ and $c^k \to \overline{c} < \infty$, then $\{f(x^k)\}$ converges to f^* R-linearly.

Proof: By Prop. 2(a), we have $f(x^k) \to f^*$, so Lemma 2 yields $f(\tilde{x}_I^k) \to f^*$ for all $I \in \mathcal{C}$, where $\tilde{x}_I^k$ is given by (9). Since X^* is compact, it follows that for all k sufficiently large we have

$$\rho(x^k; X^*) \leq \delta, \qquad \rho(\tilde{x}_I^k; X^*) \leq \delta \qquad \forall\ I \in \mathcal{C}.$$

Without loss of generality we assume that the above relation holds for all k.

(a) For each k and $I \in \mathcal{C}$, we have from Prop. 1 that $\tilde{x}_I^k$ minimizes $f(y) + \frac{1}{c^k}\Phi(y - \hat{x}_I^k)$ over y, where $\hat{x}_I^k$ is the vector in $\Re^n$ whose ith component is the ith component of x^k if $i \in I$ and, otherwise, is the ith component of $\tilde{x}_I^k$. Thus,

$$g_I^k + \frac{1}{c^k}\nabla\Phi(\tilde{x}_I^k - \hat{x}_I^k) = 0, \tag{19}$$

where g_I^k is a subgradient of f at $\tilde{x}_I^k$. Let us denote by $\overline{x}_I^k$ the element of X^* which is at minimum distance from $\tilde{x}_I^k$, that is,

$$\rho(\tilde{x}_I^k; X^*) = ||\overline{x}_I^k - \tilde{x}_I^k||.$$

From (15) with $\alpha = 1$ and using the convexity of f, we have

$$\begin{aligned}
\beta\rho(\tilde{x}_I^k; X^*) &\leq f(\tilde{x}_I^k) - f(\overline{x}_I^k) \\
&\leq \langle g_I^k, \tilde{x}_I^k - \overline{x}_I^k \rangle \\
&\leq ||g_I^k||\ ||\tilde{x}_I^k - \overline{x}_I^k|| \\
&= ||g_I^k||\ \rho(\tilde{x}_I^k; X^*).
\end{aligned}$$

Thus for all k and $I \in \mathcal{C}$ we have

$$0 \leq \big(||g_I^k|| - \beta\big)\rho(\tilde{x}_I^k; X^*). \tag{20}$$

If $c^k \to \infty$, then from (19) and the boundedness of $\{\tilde{x}_I^k - \hat{x}_I^k\}$, we have $g_I^k \to 0$ for all $I \in \mathcal{C}$, and (20) implies that for sufficiently large k, we have $\rho(\tilde{x}_I^k; X^*) = 0$ or equivalently $\tilde{x}_I^k \in X^*$, for all $I \in \mathcal{C}$. This implies that $x^{k+1} \in X^*$ for all sufficiently large k, so the algorithm terminates finitely.

If $c^k \to \overline{c} < \infty$, let $\mathcal{K}$ be an infinite subsequence of the set of positive integers such that for every $I \in \mathcal{C}$, either $\alpha_I^k > 0$ for all $k \in \mathcal{K}$ or $\alpha_I^k = 0$ for all $k \in \mathcal{K}$; since $\mathcal{C}$ is a finite set, such a subsequence exists. Let $\overline{\mathcal{C}}$ be the subset of index sets $I \in \mathcal{C}$ such that $\alpha_I^k > 0$ for all $k \in \mathcal{K}$. Using the assumption (17), we have

$$\alpha_I^k \geq \overline{\alpha} \quad \forall k \in \mathcal{K} \text{ and } I \in \overline{\mathcal{C}}, \qquad \alpha_I^k = 0 \quad \forall\, k \in \mathcal{K} \text{ and } I \notin \overline{\mathcal{C}}.$$

Then from (13) we obtain

$$\lim_{k \to \infty,\ k \in \mathcal{K}} \sum_{I \in \overline{\mathcal{C}}} \Phi\big(\tilde{x}_I^k - \hat{x}_I^k\big) = 0,$$

implying

$$\lim_{k \to \infty,\ k \in \mathcal{K}} \Phi\big(\tilde{x}_I^k - \hat{x}_I^k\big) = 0 \qquad \forall\, I \in \overline{\mathcal{C}}.$$

Therefore, we have

$$\lim_{k \to \infty,\ k \in \mathcal{K}} \nabla\Phi\big(\tilde{x}_I^k - \hat{x}_I^k\big) = 0 \qquad \forall\, I \in \overline{\mathcal{C}},$$

or equivalently, in view of (19),

$$\lim_{k \to \infty,\ k \in \mathcal{K}} g_I^k = 0 \qquad \forall\, I \in \overline{\mathcal{C}}.$$

It follows from (20) that for all $I \in \overline{\mathcal{C}}$ and all $k \in \mathcal{K}$ sufficiently large we have $\rho(\tilde{x}_I^k; X^*) = 0$. Using (10) and the fact $\alpha_I^k = 0$ for all $I \notin \overline{\mathcal{C}}$ and $k \in \mathcal{K}$, we obtain $x^{k+1} \in X^*$ for all $k \in \mathcal{K}$ sufficiently large. Since the choice of the subsequence $\mathcal{K}$ was arbitrary and a finite number of such subsequences comprise all integers beyond some index, this shows that the algorithm terminates finitely.

(b) For each k, let us denote by $\overline{x}^k$ the element of X^* which is at minimum distance from x^k, that is,

$$\rho(x^k; X^*) = ||\overline{x}^k - x^k|| \qquad \text{and} \qquad f(\overline{x}^k) = f^*.$$

We have for each $I \in \mathcal{C}$,

$$\begin{aligned} f(\tilde{x}_I^k) - f^* &\leq f(\overline{x}^k) + \frac{1}{c^k} \sum_{i \in I} \phi(\overline{x}_i^k - x_i^k) - f^* \\ &\leq \frac{1}{c^k} \Phi(\overline{x}^k - x^k) \\ &\leq \frac{M}{c^k} ||\overline{x}^k - x^k||^\gamma \\ &= \frac{M}{c^k} \rho(x^k; X^*)^\gamma \\ &\leq \frac{M}{c^k} \left(\frac{f(x^k) - f^*}{\beta} \right)^{\gamma/\alpha}, \end{aligned}$$

where the third inequality follows from (18) and the last inequality follows from (15). By using the convexity of f and (10), we obtain for all k that

$$\frac{f(x^{k+1}) - f^*}{\big(f(x^k) - f^*\big)^{\gamma/\alpha}} \leq \frac{M}{c^k \beta^{\gamma/\alpha}}. \tag{21}$$

This proves the result.

(c) Using the compactness of X^* we have that there exists $\delta' > 0$ such that $\rho(x; X^*) \leq \delta$ for all x such that $f(x) - f^* \leq \delta'$. Given $x \in \Re^n$ such that $f(x) - f^* \leq \delta'$, $c > 0$, and a subset $I \subset \{1, \ldots, n\}$, let x' and x'' be given by (4) and (5), respectively. By Prop. 1, x' satisfies (6) so

$$f(x') + \frac{1}{c}\langle x' - x'', y - x' \rangle \leq f(y) \qquad \forall\ y.$$

Let x^* be the element of X^* for which $||x' - x^*|| = \rho(x'; X^*)$. Setting $y = x^*$ in the above relation and rearranging terms, we obtain

$$f(x') - f^* \leq \frac{1}{c}\langle x'' - x', x^* - x' \rangle \leq \frac{1}{c}||x'' - x'||||x^* - x'||.$$

By Lemma 2 and the assumption $f(x) - f^* \leq \delta'$, we have $f(x') - f^* \leq \delta'$, so that $\rho(x'; X^*) \leq \delta$ and (15) with $\alpha = 2$ yields

$$\beta||x^* - x'||^2 \leq f(x') - f^* \leq \frac{1}{c}||x'' - x'||||x^* - x'||$$

or, equivalently,

$$||x^* - x'|| \leq \frac{1}{c\beta}||x'' - x'||.$$

Then, by (3) and $\Phi(\cdot) = \frac{1}{2}||\cdot||^2$,

$$F_c(x') - f^* \leq f(x^*) + \frac{1}{2c}||x^* - x'||^2 - f^* \leq \frac{1}{2c(c\beta)^2}||x'' - x'||^2.$$

On the other hand, we have from the proof of Lemma 1(b) that

$$F_c(x') \leq F_c(x'') - \frac{1}{2c}||x' - x''||^2,$$

which together with the above relation yields

$$F_c(x') - f^* \leq \frac{1}{(c\beta)^2}\big(F_c(x'') - F_c(x')\big).$$

Rearranging terms and using the fact [cf. (7)] $F_c(x'') \leq F_c(x)$, we finally obtain

$$F_c(x') - f^* \leq \frac{1}{(c\beta)^2 + 1}\big(F_c(x) - f^*\big). \tag{22}$$

Consider now a sequence $\{x^k\}$ generated by the parallel PPM algorithm with $\{c^k\}$ monotonically nondecreasing. Since X^* is compact, by Prop. 2(a), we have $f(x^k) \to f^*$ so that $f(x^k) - f^* \leq \delta'$ for

all k large enough. Fix any such k. By (9), we can apply (22) with $x = x^k$, $c = c^k$ and $x' = \tilde{x}_I^k$ for every $I \in \mathcal{C}$ and obtain

$$F_{c^k}(\tilde{x}_I^k) - f^* \leq \frac{1}{(c^k\beta)^2 + 1}\big(F_{c^k}(x^k) - f^*\big) \qquad \forall\, I \in \mathcal{C}.$$

Then, using the definition (10)-(11) of x^{k+1} and the convexity of F_{c^k}, we obtain

$$F_{c^k}(x^{k+1}) - f^* \leq \frac{1}{(c^k\beta)^2 + 1}\big(F_{c^k}(x^k) - f^*\big) \qquad \forall\, I \in \mathcal{C}.$$

Since $F_{c^{k+1}}(x^{k+1}) \leq F_{c^k}(x^{k+1})$ [cf. $c^{k+1} \geq c^k$ and (3)], this implies

$$F_{c^{k+1}}(x^{k+1}) - f^* \leq \frac{1}{(c^k\beta)^2 + 1}\big(F_{c^k}(x^k) - f^*\big).$$

We have that $\{c^k\}$ is bounded below by c^0, so it follows that $\big\{F_{c^k}(x^k)\big\}$ converges linearly. Since $f(x^{k+1}) \leq F_{c^k}(x^k)$ for all k, we obtain that $\big\{f(x^k)\big\}$ converges R-linearly and part (d) is proven. **Q.E.D.**

The preceding proof of part (b) also shows that if $\alpha = \gamma$ and $c^k \to \bar{c} \in (\beta/M, \infty)$, then $\big\{f(x^k)\big\}$ converges linearly [see (21)].

The preceding analysis assumes that the set of minimizers of f is bounded. We show below that this assumption can be removed if the minimized function f is differentiable on its effective domain and has a growth property similar to that given by (15) with $\alpha = 2$. This result will be useful when we analyze dual applications of the partial proximal algorithm in Section 5, for which the set of minimizers of f is frequently unbounded (see Prop. 7).

Assumption B: The set of minimizers of f, denoted X^*, is nonempty and f has the special form:

$$f(x) = \begin{cases} g(x) & \text{if } x \in C \\ \infty & \text{otherwise,} \end{cases} \tag{23}$$

where C is a nonempty closed convex set in $\Re^n$ and $g : \Re^n \mapsto \Re$ is a convex differentiable function. Furthermore, there exist scalars $\beta > 0$ and $\delta > 0$ such that

$$\beta\rho(x; X^*) \leq ||x - P_C[x - \nabla g(x)]|| \qquad \forall\, x \in C \text{ with } ||x - P_C[x - \nabla g(x)]|| \leq \delta, \tag{24}$$

where $P_C[\cdot]$ denotes the orthogonal projection onto C and $\rho(x; X^*)$ is the distance from x to X^* defined by (16).

The growth condition (24) differs from the growth condition (15) (with $\alpha = 2$) mainly in that the cost difference $f(x) - f^*$ is approximated by the norm of a certain residual function squared. This difference is nonetheless significant for it turns out that the partial proximal algorithm drives

the latter to zero even when X^* is unbounded (see the proof below). In general, verifying that the condition (24) holds is not easy. However, it is known that this condition holds when g is strongly convex [Pan87] or when C is a polyhedral set and g is the composition of a strongly convex function with an affine mapping [LuT92]. (See [LuT91] and Section 5 for additional discussions of this condition.)

Proposition 4: Let Assumption B hold and let $\{x^k\}$ be a sequence generated by the parallel PPM algorithm (9)-(11) with $\{c^k\}$ monotonically nondecreasing and with $\phi(\cdot) = \frac{1}{2}|\cdot|^2$. Then $\{f(x^k)\}$ converges to $\min_x f(x)$ R-linearly.

Proof: First, we show that, for any $c > 0$, the function F_c inherits from f a property similar to the growth condition (24). Fix any $x \in \Re^n$ and let

$$\tilde{x} = \arg\min_{y \in \Re^n} \left\{ f(y) + \frac{1}{2c}||y - x||^2 \right\}.$$

Then, by $\Phi(\cdot) = \frac{1}{2}||\cdot||^2$, we have $x - \tilde{x} = c\nabla F_c(x)$ and, by using (23), we also have that $\tilde{x}$ is a minimizer of the function $y \mapsto g(y) + \frac{1}{2c}||y - x||^2$ over C, so that

$$\tilde{x} = P_C[x - c\nabla g(\tilde{x})].$$

Thus,

$$||\tilde{x} - P_C[\tilde{x} - c\nabla g(\tilde{x})]|| = ||P_C[x - c\nabla g(\tilde{x})] - P_C[\tilde{x} - c\nabla g(\tilde{x})]|| \leq ||x - \tilde{x}||,$$

where the second inequality follows from the nonexpansive property of the projection operator $P_C[\cdot]$. Since $||\tilde{x} - P_C[\tilde{x} - cd]|| \geq c||\tilde{x} - P_C[\tilde{x} - d]||$ for any $d \in \Re^n$ (see Lemma 1 in [GaB84]), this implies

$$c||\tilde{x} - P_C[\tilde{x} - \nabla g(\tilde{x})]|| \leq ||x - \tilde{x}||.$$

Then, it readily follows from (24) that

$$c\beta\rho(\tilde{x}; X^*) \leq ||x - \tilde{x}||, \qquad \text{whenever } ||x - \tilde{x}|| \leq c\delta.$$

Since $x - \tilde{x} = c\nabla F_c(x)$ and, by the triangle inequality, $\rho(x; X^*) \leq ||x - \tilde{x}|| + \rho(\tilde{x}; X^*)$, this shows that

$$\beta\rho(x; X^*) \leq (c\beta + 1)||\nabla F_c(x)||, \qquad \text{whenever } ||\nabla F_c(x)|| \leq \delta.$$

Also, denoting $f^* = \min_y f(y)$, we have from (3) and $\Phi(\cdot) = \frac{1}{2}||\cdot||^2$ that

$$F_c(x) = f(\tilde{x}) + \frac{1}{2c}||\tilde{x} - x||^2 \leq f^* + \frac{1}{2c}||x^* - x||^2 \qquad \forall\, x^* \in X^*,$$

so that $F_c(x) \leq f^* + \frac{1}{2c}\rho(x; X^*)^2$. Combining this with the previous relation yields

$$F_c(x) \leq f^* + \frac{1}{2c}(c + 1/\beta)^2||\nabla F_c(x)||^2 \qquad \text{whenever } ||\nabla F_c(x)|| \leq \delta. \tag{25}$$

Consider now a sequence $\{x^k\}$ generated by the parallel PPM algorithm with $\{c^k\}$ monotonically nondecreasing. By Prop. 2(b), we have $\nabla F_{c^k}(x^k) \to 0$, so that $||\nabla F_{c^k}(x^k)|| \leq \delta$ for all k sufficiently large. For any such k we have from (25) that

$$F_{c^k}(x^k) \leq f^* + \frac{1}{2c^k}(c^k + 1/\beta)^2 ||\nabla F_{c^k}(x^k)||^2.$$

In addition, by following the proof of Prop. 2(b), we see that

$$F_{c^{k+1}}(x^{k+1}) - F_{c^k}(x^k) \leq -\frac{c^k}{4}||\nabla F_{c^k}(x^k)||^2.$$

Combining the above two relations and rearranging terms yields

$$F_{c^{k+1}}(x^{k+1}) - f^* \leq \left(1 - \frac{(c^k\beta)^2}{2(c^k\beta+1)^2}\right)(F_{c^k}(x^k) - f^*).$$

We have that $\{c^k\}$ is bounded below by c^0, so it follows that $\{F_{c^k}(x^k)\}$ converges linearly to f^*. Since $f(x^{k+1}) \leq F_{c^k}(x^k)$ for all k, we obtain that $\{f(x^k)\}$ converges R-linearly to f^*. **Q.E.D.**

5. RELATION TO MULTIPLIER METHODS

We assume throughout this section that $\phi(\cdot) = \frac{1}{2}|\cdot|^2$ [so that $\Phi(\cdot)$ is the quadratic function $\frac{1}{2}||\cdot||^2$], and we show that partial proximal iterations correspond to augmented Lagrangian iterations with partial elimination of constraints. This indicates a possible application area of the parallel PPM algorithm of Section 3 and establishes its relation to the constraint distribution method of [FeM91]. In addition, by applying the convergence results of Section 4, we analyze the rate of convergence of these augmented Lagangian iterations under much weaker assumptions than those given in [FeM91]. For example, we establish linear rate of convergence for the dual cost of the multipliers, assuming that the constraint functions are affine, and the cost function is the sum of the indicator function for a polyhedral set and a strongly convex differentiable function with Lipschitz continuous gradient (see Prop. 7). In contrast, the linear rate of convergence result in [FeM91] in addition assumes that the cost function is quadratic. (On the other hand, the analysis of [FeM91] establishes the stronger result of linear rate of convergence for the multipliers.)

Consider the following convex program

$$\text{minimize} \quad h_0(z) \tag{26a}$$

$$\text{subject to} \quad h_1(z) \leq 0, \cdots, h_n(z) \leq 0, \tag{26b}$$

where $h_0, \ldots, h_n$ are closed proper convex functions in $\Re^m$ ($m \geq 1$). We can also allow for linear equality constraints in the above problem but, for simplicity, we will not consider this more general case.

For any convex function g, we denote by dom g the effective domain of g, i.e., dom $g = \{ z \mid g(z) < +\infty \}$. For any convex set C, we denote by int(C) and ri(C), respectively, the interior and the relative interior of C. We make the following standing assumptions regarding the convex program (26):

Assumption C:

(a) There exists a $\bar{z} \in$ ri(dom h_0) satisfying $h_i(\bar{z}) \leq 0$ for all i, with strict inequality holding whenever h_i is not affine.

(b) The level sets of the program (26), namely, sets of the form

$$\{ z \mid h_0(z) \leq \xi, h_1(z) \leq 0, \ldots, h_n(z) \leq 0 \}$$

with $\xi \in \Re$, are bounded.

Note that by part (a) of Assumption C, the program (26) has at least one feasible solution. This, together with part (b) of Assumption C, implies that the set of optimal solutions for (26) is nonempty and compact.

By associating a Lagrange multiplier x_i with the constraint $h_i(z) \leq 0$ for every i, we obtain the following dual function:

$$f(x) = \begin{cases} \sup_z \{-\langle x, h(z)\rangle - h_0(z)\} & \text{if } x \geq 0, \\ +\infty & \text{otherwise,} \end{cases} \tag{27}$$

where we denote by $h(z)$ the vector in $\Re^n$ whose ith component is $h_i(z)$ and by x the vector in $\Re^n$ whose ith component is x_i. It is well known that f is a closed proper convex function.

It is known that when Assumption C holds, the set of Kuhn-Tucker vectors for the convex program (26) is nonempty and equals the set of minimizers of f (see [Roc70, Th. 28.2]). Moreover, strong duality holds in the sense that the optimal value of problem (26) equals the negative of the minimum value of f. Thus, we can consider solving problem (26) by minimizing the dual function f of (27) and, for this purpose, we can use the parallel PPM algorithm (9)-(11). We show below that, for $\phi(\cdot) = \frac{1}{2}|\cdot|^2$, the proximal minimization step (9) in this algorithm is well defined and can be implemented with the use of quadratic augmented Lagrangian functions. Fix any nonempty subset I of $\{1, \ldots, n\}$, any $x \in \Re^n$ and any scalar $c > 0$. Consider the following convex program associated with I, x and c:

$$\text{minimize} \quad h_0(z) + \frac{1}{2c} \sum_{i \in I} [x_i + c h_i(z)]_+^2 \tag{28a}$$

$$\text{subject to} \quad h_i(z) \leq 0, \ i \notin I, \tag{28b}$$

where, for any number a, we denote by $[a]_+$ the positive part of a, i.e., $[a]_+ = \max\{0, a\}$. This program has at least one feasible solution (namely, $\bar{z}$) and its level sets are bounded [since any direction of unboundedness for this program would also be a direction of unboundedness for the program (26)], so it has at least one optimal solution. Let z' be any such optimal solution. Notice that the program (28) has a feasible solution (namely $\bar{z}$), which is in the relative interior of the effective domain of its cost function and satisfies with strict inequality all constraints for which h_i is not affine. Then, by Th. 28.2 in [Roc70], the program (28) has a Kuhn-Tucker vector. Fix any such Kuhn-Tucker vector and let x_i', $i \notin I$, denote its component associated with the constraint $h_i(z) \leq 0$. Let x' be the vector in $\Re^n$ whose ith component is x_i' for all $i \notin I$ and, otherwise, is

$$x_i' = [x_i + ch_i(z')]_+ \qquad \forall\, i \in I. \tag{29}$$

We claim that x' is a minimizer of the function

$$y \mapsto f(y) + \frac{1}{2c} \sum_{i \in I} |y_i - x_i|^2. \tag{30}$$

To see this, notice from the Kuhn-Tucker conditions for the program (28) that

$$x_i' = [x_i' + ch_i(z')]_+ \qquad \forall\, i \notin I, \tag{31}$$

and

$$0 \in \partial h_0(z') + \sum_{i \in I} [x_i + ch_i(z')]_+ \partial h_i(z') + \sum_{i \notin I} x_i' \partial h_i(z').$$

This equation together with (29) yields

$$0 \in \partial h_0(z') + \sum_{i=1}^{n} x_i' \partial h_i(z'),$$

implying

$$z' = \arg\min_z \left\{ \langle x', h(z) \rangle + h_0(z) \right\}. \tag{32}$$

Let us write (29) and (31) equivalently as

$$0 \in T_i - h_i(z') + \frac{x_i' - x_i}{c} \quad \forall\, i \in I, \qquad 0 \in T_i - h_i(z') \quad \forall\, i \notin I,$$

where T_i is the interval $[0, +\infty)$ if $x_i' = 0$ and otherwise is just the origin $\{0\}$. From (32) and the definition of f [cf. (27)], we see that the Cartesian product $(T_1 - h_1(z')) \times \cdots \times (T_n - h_n(z'))$ is precisely $\partial f(x')$. This together with the above relation shows that x' is a minimizer of the function given by (30).

The above discussion shows that the parallel PPM algorithm with quadratic proximal term, applied to minimizing the dual function f of (27), is well defined and that each partial proximal

minimization (9) can be achieved by solving a convex program of the form (28). A key feature of the program (28) is that only a subset of the constraints are eliminated. By carefully choosing the subsets to eliminate, one can preserve special structures of the cost function and perhaps also attain a faster rate of convergence; see [Ber82, Section 2.4], [Dun89], and [Alj90] for discussions of augmented Lagrangian methods of this type.

The above dual application of the parallel PPM algorithm is closely related to the parallel constraint distribution algorithm of [FeM91]. In particular, by noting that the program (28) is identical in form to that appearing in Th. 3.2 of [FeM91], we see that the two algorithms differ only in that the latter requires the subsets I effectively used at each iteration to form a partition of $\{1,\ldots,n\}$ and that, instead of taking a convex combination of the $\tilde{x}_I^k$'s, it extracts the coordinates indexed by I from $\tilde{x}_I^k$ to form x^{k+1}. Thus, the parallel PPM algorithm updates in a manner reminiscent of Cimmino's method [Cim38], while the parallel constraint distribution algorithm updates in a manner reminiscent of a Jacobi method. The subsequent paper [Fer91a] uses updates similar to the ones of the present paper and presents computational results showing an improved performance over the algorithm of [FeM91].

Under a strong regularity assumption that guarantees boundedness of the set of Kuhn-Tucker vectors for the convex program (26), we immediately obtain as a consequence of Prop. 2(a) the following convergence result for the parallel PPM algorithm.

Proposition 5: Assume that there is a point in dom h_0 satisfying all the constraints in (26b) with strict inequality. Consider the parallel PPM algorithm (9)-(11) with $\{c^k\}$ monotonically nondecreasing and with $\phi(\cdot) = \frac{1}{2}|\cdot|^2$, applied to minimize f given by (27). Then a sequence $\{x^k\}$ generated by the algorithm is bounded, each of its cluster points is a minimizer of f, and $\{-f(x^k)\}$ converges to the optimal value of the program (26).

Proof: By the given hypothesis, the convex program (26) is strictly consistent in the terminology of [Roc70, p. 300]. Since the optimal value of problem (26) is finite, it follows from Corollary 29.1.5 in [Roc70] that the Kuhn-Tucker vectors for problem (26) form a nonempty compact convex subset of $\Re^n$. Since these Kuhn-Tucker vectors are precisely the minimizers of f, the hypothesis of Prop. 2(a) holds, and the result follows from that proposition. **Q.E.D.**

By translating the growth conditions (15) and (24) on f into conditions on h_0 and $h_1, ..., h_m$, and then applying Props. 3 and 4, we analogously obtain the following two rate of convergence results for the parallel PPM algorithm.

Proposition 6: Assume that there is a point in dom h_0 satisfying all the constraints in (26b) with strict inequality. Furthermore, assume that there exist scalars $\alpha > 1$, $\beta > 0$, and $\delta > 0$ such

that

$$p(u) - p(0) - \langle u, x^* \rangle \leq (\alpha - 1)\beta \left(\frac{||u||}{\alpha\beta} \right)^{\frac{\alpha}{\alpha-1}} \quad \forall\, x^* \in \partial p(0) \text{ and } u \text{ with } ||u|| \leq \alpha\beta\delta^{\alpha-1},$$

where $p : \Re^n \to (-\infty, +\infty]$ is the perturbation function given by $p(u) = \min\{\ h_0(z) \mid h(z) \leq u\ \}$. Let $\{x^k\}$ be a sequence generated by the parallel PPM algorithm (9)-(11) with $\{c^k\}$ monotonically nondecreasing and with $\phi(\cdot) = \frac{1}{2}|\cdot|^2$, applied to minimize f given by (27). Let $f^* = \min_x f(x) = -p(0)$.

(a) If $1 < \alpha < 2$, then $\{f(x^k)\}$ converges to f^* superlinearly with order $2/\alpha$.

(b) If $\alpha = 2$ and $c^k \to \infty$, then $\{f(x^k)\}$ converges to f^* superlinearly.

(c) If $\alpha = 2$ and $c^k \to \bar{c} < \infty$, then $\{f(x^k)\}$ converges to f^* R-linearly.

Proof: As was shown in the proof of Prop. 5, the set X^* is nonempty and compact. We show below that f satisfies (15) with α, β, δ as given, so the claim immediately follows from Prop. 3(b)-(c).

Fix any $x \in \Re^n$ with $\rho(x; X^*) \leq \delta$. First assume that $x \notin X^*$. Let x^* be the minimizer of f nearest to x, i.e., $\rho(x; X^*) = ||x - x^*||$. Also let

$$u = \alpha\beta(x - x^*)||x - x^*||^{\alpha-2}.$$

It is well known that p is the conjugate function of f, so that, by [Roc70, Thm. 23.5],

$$x^* \in \partial p(0).$$

Also, direct calculation finds that $||u|| \leq \alpha\beta\delta^{\alpha-1}$ and

$$(\alpha - 1)\beta \left(\frac{||u||}{\alpha\beta} \right)^{\frac{\alpha}{\alpha-1}} + \langle u, x^* \rangle = \langle u, x \rangle - \beta||x - x^*||^\alpha.$$

Thus, the hypothesis on p yields

$$p(u) - p(0) \leq \langle u, x \rangle - \beta||x - x^*||^\alpha.$$

Also, we have

$$p(u) = \sup_y \{\langle u, y \rangle - f(y)\} \geq \langle u, x \rangle - f(x), \qquad p(0) = -f^*,$$

which together with the above inequality yields

$$-f(x) + f^* \leq -\beta||x - x^*||^\alpha.$$

Rearranging terms and using $\rho(x; X^*) = ||x - x^*||$, we obtain

$$\beta\rho(x; X^*)^\alpha \leq f(x) - f^*.$$

Second, assume that $x \in X^*$. Then the above relation holds trivially. **Q.E.D.**

As is shown by the preceding proof, the growth condition in Prop. 6 can alternatively be replaced by the growth condition (15) on the dual function f. Depending on the problem structure, one condition may be easier to verify than the other.

Proposition 7: Assume that the cost function h_0 is the sum of the indicator function of a polyhedral set and a strongly convex differentiable function whose gradient is Lipschitz continuous everywhere. Also assume that the constraint functions $h_1, ..., h_n$ are affine. Let $\{x^k\}$ be a sequence generated by the parallel PPM algorithm (9)-(11) with $\{c^k\}$ monotonically nondecreasing and with $\phi(\cdot) = \frac{1}{2}|\cdot|^2$, applied to minimize the dual function f given by (27). Then $\{-f(x^k)\}$ converges R-linearly to the optimal value of the convex program (26).

Proof: It can be seen that, in this case, the convex program (26) is a special case of the convex program (2.2) studied in [LuT91] and that Assumptions A and B therein hold. Then, by Thm. 4.1 in [LuT91], f satisfies Assumption B when restricted to the level set $\{ x \mid x \geq 0,\ f(x) \leq f(x^0) \}$. Since, by Prop. 2(b), $\{f(x^k)\}$ is monotonically nonincreasing so the sequence $\{x^k\}$ lies in this level set, we can invoke Prop. 4 to conclude that $\{f(x^k)\}$ converges to $\min_x f(x)$ R-linearly. **Q.E.D.**

We remark that Prop. 7 is similar to the rate of convergence results obtained in [FeM91] and [Fer91a], but these references treat only the case where h_0 is strongly convex quadratic and $h_1, ..., h_n$ are affine.

REFERENCES

[Alj90] Aljazzaf, M., Multiplier Methods with Partial Elimination of Constraints for Nonlinear Programming, Ph.D. Thesis, Department of Mathematics, North Carolina State University, Raleigh, NC, 1990.

[BeT89] Bertsekas, D. P., and Tsitsiklis, J. N., *Parallel and Distributed Computation: Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ, 1989.

[Ber75] Bertsekas, D. P., Necessary and Sufficient Conditions for a Penalty Method to be Exact, *Proceedings of the Symposium on Large-Scale Systems*, Udine, Italy, 1976, pp. 353-359.

[Ber82] Bertsekas, D. P., *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York, NY, 1982.

[CeZ92] Censor, Y., and Zenios, S. A., Proximal Minimization Algorithm with D-Functions, *Journal of Optimization Theory and Applications*, Vol. 73, 1992, pp. 451-463.

[ChT90] Chen, G., and Teboulle, M., Convergence Analysis of a Proximal-Like Minimization Algorithm Using Bregman Functions, Department of Mathematics and Statistics Report, University of Maryland, Baltimore County, MD, 1990; to appear in *SIAM Journal on Optimization.*

[Cim38] Cimmino, G., Calcolo Approssimato per le Soluzioni dei Sistemi di Equazioni Lineari, *Ricerca Scientifica*, XVI Seria II, Anno IX, Vol. 1, 1938, pp. 326-333.

[Dun89] Dunn, J. C., Formal Augmented Newtonian Projection Methods for Continuous-Time Optimal Control, Proceedings of 28th IEEE Conference on Decision and Control, Tampa, FL, 1989.

[EcB92] Eckstein, J., and Bertsekas, D. P., On the Douglas-Rachford Splitting Method and the Proximal Point Algorithm for Maximal Monotone Operators, *Mathematical Programming*, Vol. 55, 1992, pp. 263-275.

[Eck93] Eckstein, J., Nonlinear Proximal Point Algorithms using Bregman Functions, with Applications to Convex Programming, *Mathematics of Operations Research*, Vol. 18, 1993, pp. 202-226.

[FeM91] Ferris, M. C., and Mangasarian, O. L., Parallel Constraint Distribution, *SIAM Journal on Optimization*, Vol. 1, 1991, pp. 487-500.

[Fer91a] Ferris, M. C., Parallel Constraint Distribution in Convex Quadratic Programming, Department of Computer Science Report, University of Wisconsin, Madison, WI, 1991.

[Fer91b] Ferris, M. C., Finite Termination of the Proximal Point Algorithm, *Mathematical Programming*, Vol. 50, 1991, pp. 359-366.

[GaB84] Gafni, E. M., and Bertsekas, D. P., Two-Metric Projection Methods for Constrained Optimization, *SIAM Journal on Control and Optimization*, Vol. 22, 1984, pp. 936-964.

[GoT79] Golshtein, E. G., and Tretjiakov, N. V., Modified Lagrangians in Convex Programming and their Generalizations, *Mathematical Programming Studies*, Vol. 10, 1979, pp. 86-97.

[Gul91] Güler, O., On the Convergence of the Proximal Point Algorithm for Convex Minimization, *SIAM Journal on Control and Optimization*, Vol. 29, 1991, pp. 403-419.

[Ha90] Ha, C. D., A Generalization of the Proximal Point Algorithm, *SIAM Journal on Control and Optimization*, Vol. 28, 1990, pp. 503-512.

[BoB72] Kort, B. W., and Bertsekas, D. P., A New Penalty Function Method for Constrained Minimization, Proc. 1971 IEEE Decision and Control Conference, New Orleans, LA, December 1972.

[KoB76] Kort, B. W., and Bertsekas, D. P., Combined Primal-Dual and Penalty Methods for Convex Programming, *SIAM Journal on Control and Optimization*, Vol. 14, 1976, pp. 268-294.

[Luq84] Luque, F. J., *Nonlinear Proximal Point Algorithms*, Ph.D. Thesis, Department of Civil Engineering and Operations Research Center, Massachusetts Institute of Technology, Cambridge, MA, 1984.

[Luq86] Luque, F. J., The Nonlinear Proximal Point Algorithm, Laboratory for Information and Decision Systems Report P-1598, Massachusetts Institute of Technology, Cambridge, MA, 1986.

[LuT91] Luo, Z.-Q., and Tseng, P., On the Convergence Rate of Dual Ascent Methods for Strictly Convex Minimization, Communications Research Laboratory, McMaster University, Hamilton, On., and Department of Mathematics, University of Washington, Seattle, WA, 1991; to appear in *Mathematics of Operations Research*.

[LuT92] Luo, Z.-Q., and Tseng, P., On the Linear Convergence of Descent Methods for Convex Essentially Smooth Minimization, *SIAM Journal on Control and Optimization*, Vol. 30, 1992, pp. 408-425.

[Mar70] Martinet, B., Regularisation d'inéquations variationelles par approximations successives, *Revue Française d'Automatique et Informatique Recherche Opérationelle*, Vol. 4, 1970, pp. 154-159.

[Mar72] Martinet, B., Determination approchée d'un point fixe d'une application pseudo-contractante, *Compte Rendu Académie des Sciences de Paris*, Vol. 274, 1972, pp. 163-165.

[OrR70] Ortega, J. M., and Rheinboldt, W. C., Iterative Solution of Nonlinear Equations in Several Variables, Academic Press, New York, NY, 1970.

[Pan87] Pang, J.-S., A Posteriori Error Bounds for the Linearly-Constrained Variational Ineqality Problem, *Mathematics of Operations Research*, Vol. 12, 1987, pp. 474-484.

[Roc70] Rockafellar, R. T., *Convex Analysis*, Princeton University Press, Princeton, NJ, 1970.

[Roc73] Rockafellar, R. T., A Dual Approach to Solving Nonlinear Programming Problems by Unconstrained Minimization, *Mathematical Programming*, Vol. 5, 1973, pp. 354-373.

[Roc76] Rockafellar, R. T., Monotone Operators and the Proximal Point Algorithm, *SIAM Journal on Control and Optimization*, Vol. 14, 1976, pp. 877-898.

[TsB90] Tseng, P., and Bertsekas, D. P., On the Convergence of the Exponential Multiplier Method for Convex Programming, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology Report P-1995, Cambridge, MA, 1990; to appear in *Mathematical Programming*.